

APPLICATION FOR
UNITED STATES PATENT
IN THE NAME OF

Jeffrey Skolnick, Mariusz Milik, and Andrzej Kolinski

of

The Scripps Research Institute

FOR

**Prediction of Relative Binding Motifs of Biologically Active
Peptides and Peptide Mimetics**

John Land
FISH & RICHARDSON
4225 Executive Square, Suite 1400
La Jolla, CA 92037
(619) 678-5070 voice
(619) 678-5099 fax

Date of Deposit: 5/23/97
I hereby certify under 37 CFR 1.10 that this correspondence is being deposited with the United States Postal Service as "Express Mail Post Office To Addressee" with sufficient postage on the date indicated above and is addressed to the Commissioner of Patents and Trademarks, Washington, D.C. 20231.

Christopher Horne
Christopher Horne

DOCKET NO. 07300/034001

EXPRESS MAIL NO. EM12276061545

PREDICTION OF RELATIVE BINDING MOTIFS OF BIOLOGICALLY ACTIVE PEPTIDES AND PEPTIDE MIMETICS

BACKGROUND OF THE INVENTION

1. *Field of the Invention*

5 This invention relates to computer-assisted analysis of biological molecules, particularly of biologically active peptides and peptide mimetics.

2. *Description of Related Art*

10 With the ever increasing plethora of biological information, the new branch of biological sciences called bioinformatics has become increasingly important. Bioinformatics seeks to translate the mass of protein (polypeptide) sequence information into knowledge of structure and more importantly, function.

15 One category of peptides where structure and function information would be useful are Class I major histocompatibility complex (MHC) molecules (in humans, the MHC is called HLA). MHC molecules are cell surface proteins that present bound peptides. These peptides are analyzed by immuno-surveillant cytotoxic T-cells (CTLs) to identify foreign or unhealthy cells for removal. Understanding this process is important, as it constitutes the primary immunological defense against viruses and perhaps tumor causing cells. It is also a major component responsible for transplant rejection. A. Townsend and H. Bodmer, *Annu. Rev. Immunol.* 7, 601 (1989); J.W. Yewdell and J.R. Binnink, *Adv. Immunol.* 52, 1 (1992). Since the affinity of the bound peptides largely determines the stability of the expressed class I molecules and their recognition by CTLs, it is crucial to determine the rules of peptide binding by class I molecules.

2
20250-04225000

Analyses of peptides eluted from class I MHC molecules reveal that they are short, usually 8-10 amino acids long, with particular amino acids occurring in specific, anchor positions with a very high frequency. Highly conserved pockets accommodate these anchor amino acids as well as the peptide amino and carboxy termini. The carboxy terminal pocket is considerably less constraining than the amino terminus (M. Matsumura, Y. Saito, M.R. Jackson, E.S. Song and P.A. Peterson, *J. Biol. Chem.* 267(33), 23589 (1992); E.J. Collins, E.N. Garboczi and D.C. Wiley, *Nature* 371, 629 (1994)), suggesting the possibility of using a phage display analysis for peptide screening.

Binding analyses with synthetic peptides have confirmed the importance of the anchor residues but have also revealed amino acid preferences at other positions. These secondary anchor residues can have profound effects on binding affinities, as peptide binding to human class I molecules can vary by over four orders of magnitude. Furthermore, combinations of anchor amino acids are restricted, making the binding rules complex. Hence predictions based solely on anchor amino acids are at best about 20% accurate. J. Ruppert, J. Sidney, E. Celis, R. T. Kubo, H.M. Grey and A. Sette, *Cell* 74, 929 (1993). It would be desirable to have an analysis ~~is required~~ that tests a large number of peptide sequences and considers the correlated effects of amino acids.

Artificial intelligence and pattern recognition methods may prove to be powerful tools in the bioinformatics field. For example, an artificial neural network (ANN) has been successfully applied to predict mitochondrial precursor cleavage sites (G. Schneider, J. Schuchhardt and P. Wrede, *Biophys. J.* 68, 434 (1995)) and membrane-spanning amino acid sequences (R. Lohmann, G. Schanider, D. Behrens and P. Wrede, *Protein Science* 3, 1597 (1994); M. Milik and J. Skolnick, in: "Proceedings of Fourth Annual Conference on Evolutionary Programming", MIT Press, La Jolla (1995)). However, to date, ANN analysis has not been successfully applied to prediction of binding motifs of biologically active peptides and peptide mimetics. The present invention provides a method and system for accomplishing this goal.

SUMMARY OF THE INVENTION

The invention comprises a general neural network based method and system for identifying relative peptide binding motifs from limited experimental data. In particular, an artificial neural network (ANN) is trained with peptides with known sequence and function (*i.e.*, binding strength) identified from a phage display library. The ANN is then challenged with unknown peptides, and predicts relative binding motifs. Analysis of the unknown peptides validate the predictive capability of the ANN.

In one example, the training peptides bind to mouse MHC class I molecule H2-K^b. Blind testing (*e.g.*, on chicken ovalbumin) correctly identified strongly binding peptides, and their relative binding strengths, in 5 of the 7 top scoring predictions from the test procedure. Upon validation analysis, the top scoring peptide was the known immunodominant peptide. Further, the second best binding peptide, since it lacked characteristic anchor residues, would have been missed using standard statistical approaches. The ability to predict antigens that bind MHC represents a significant advance in the development of vaccines and T-cell based therapeutics.

The details of the preferred embodiment of the present invention are set forth in the accompanying drawings and the description below. Once the details of the invention are known, numerous additional innovations and changes will become obvious to one skilled in the art.

BRIEF DESCRIPTION OF THE DRAWINGS

FIGURE 1 is a schematic view of the preferred peptide sequence coding scheme and the ANN architecture of the invention.

5 FIGURE 2 is a graph showing performance of the ANN on the training and testing sets
as a function of training time, measured by the number of times the whole training set was
presented to the network (epochs).

FIGURE 3 is a graph showing a competition binding assay.

Like reference numbers and designations in the various drawings indicate like elements.

0 2 3 6 2 4 9 2 . 0 5 2 3 6 7

DETAILED DESCRIPTION OF THE INVENTION

Throughout this description, the preferred embodiment and examples shown should be considered as exemplars, rather than as limitations on the present invention.

Introduction

5 The invention will be described using an example of an artificial neural network (ANN) system used to predict relative binding motifs of peptides that bind to MHC class I molecules. However, the process is general and can be applied to any peptide system. An important aspect of the present invention is the inclusion of both experimental and theoretical aspects of the problem into one, coherent procedure. Preliminary results from 10 the ANN analysis improved the interpretation of results from phage display experiments, and later experimental methods were used in blind tests of the ANN classification scheme.

Artificial Neural Networks

15 Artificial neural networks can be used to recognize patterns and “signatures” in data streams. An ANN differs from other signal processing algorithms in that it does not assume any underlying model. Rather, an ANN “learns” to detect patterns by generating a model in response to input test data having known patterns, features, or other characteristics of interest in classifying the input data. An ANN can be trained relatively easy and repeatably. Because an ANN learns to detect patterns or correlations, ANNs are very flexible and adaptable to a wide variety of situations and conditions. This flexibility and adaptability gives artificial neural networks a significant advantage over other data classification techniques. For further information on the architecture and training of multi-layer perceptron (MLP) adaptive artificial neural networks, see “Progress in Supervised Neural Networks” by Don Hush and Bill Horne, published in *IEEE Signal Processing* (January 1993).

20

25

FIGURE 1 is a schematic view of the preferred peptide sequence coding scheme and the ANN architecture of the invention. Shown is a standard multi-layer perceptron ANN 1 trained by back-propagation (BP) of error. D. Rumelhart, J. McClelland and the PDP Research Group, "Parallel Distributed Processing", MIT Press, Cambridge (1986). The 5 ANN 1 includes an input layer 2 comprising a plurality of input units 3, a hidden layer 4 comprising a plurality of hidden units 5, and an output layer 6 comprising a plurality of output units 7. In the preferred embodiment, the number of output units is two, denoted 7a and 7b. Each unit 3, 5, 7 is a processing element or "neuron", coupled by connections having adjustable numeric weights or connection strengths by which earlier layers 10 influence later ones to determine the network output.

Prior to using the ANN 1 to classify actual input data, the parameters of the ANN 1 are 15 adjusted by applying pre-characterized training data to the ANN 1. That is, training data is selected such that particular features are known to present or known to be absent. In the invention, such data comprises an appropriately coded set of input patterns (i.e., known peptide sequences having known binding affinities). See below for a discussion of the preferred coding.

Phage Display

In order to obtain training data for an ANN, a study was initiated with a peptide phage display binding analysis of the mouse MHC class I molecule K^b. Soluble K^b was purified 20 from transfected Drosophila cells. Phage display analysis has been used previously to identify MHC class II molecule binding peptides. J. Hammer, B. Takacs and F. Sinigaglia, *J. Exp. Med.* 176, 1007 (1992). Phage display libraries were obtained from Dr. G.P. Smith of ~~the~~ and the analyses were performed essentially as described in the art (S.F. Parmeley, and G.P. Smith, *Gene* 73, 305 (1988); J.K. Scott and G. P. Smith, 25 *Science* 249, 386 (1990); G.P. Smith personal communication). From the phage display, the sequences of 181 K^b binding peptides and their relative binding affinities were obtained along with the sequences of 129 non-binding sequences.

110
115
120
125
130
135
140
145
150
155
160
165
170
175
180
185
190
195
200
205
210
215
220
225
230
235
240
245
250
255
260
265
270
275
280
285
290
295
300
305
310
315
320
325
330
335
340
345
350
355
360
365
370
375
380
385
390
395
400
405
410
415
420
425
430
435
440
445
450
455
460
465
470
475
480
485
490
495
500
505
510
515
520
525
530
535
540
545
550
555
560
565
570
575
580
585
590
595
600
605
610
615
620
625
630
635
640
645
650
655
660
665
670
675
680
685
690
695
700
705
710
715
720
725
730
735
740
745
750
755
760
765
770
775
780
785
790
795
800
805
810
815
820
825
830
835
840
845
850
855
860
865
870
875
880
885
890
895
900
905
910
915
920
925
930
935
940
945
950
955
960
965
970
975
980
985
990
995
1000
1005
1010
1015
1020
1025
1030
1035
1040
1045
1050
1055
1060
1065
1070
1075
1080
1085
1090
1095
1100
1105
1110
1115
1120
1125
1130
1135
1140
1145
1150
1155
1160
1165
1170
1175
1180
1185
1190
1195
1200
1205
1210
1215
1220
1225
1230
1235
1240
1245
1250
1255
1260
1265
1270
1275
1280
1285
1290
1295
1300
1305
1310
1315
1320
1325
1330
1335
1340
1345
1350
1355
1360
1365
1370
1375
1380
1385
1390
1395
1400
1405
1410
1415
1420
1425
1430
1435
1440
1445
1450
1455
1460
1465
1470
1475
1480
1485
1490
1495
1500
1505
1510
1515
1520
1525
1530
1535
1540
1545
1550
1555
1560
1565
1570
1575
1580
1585
1590
1595
1600
1605
1610
1615
1620
1625
1630
1635
1640
1645
1650
1655
1660
1665
1670
1675
1680
1685
1690
1695
1700
1705
1710
1715
1720
1725
1730
1735
1740
1745
1750
1755
1760
1765
1770
1775
1780
1785
1790
1795
1800
1805
1810
1815
1820
1825
1830
1835
1840
1845
1850
1855
1860
1865
1870
1875
1880
1885
1890
1895
1900
1905
1910
1915
1920
1925
1930
1935
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2025
2030
2035
2040
2045
2050
2055
2060
2065
2070
2075
2080
2085
2090
2095
2100
2105
2110
2115
2120
2125
2130
2135
2140
2145
2150
2155
2160
2165
2170
2175
2180
2185
2190
2195
2200
2205
2210
2215
2220
2225
2230
2235
2240
2245
2250
2255
2260
2265
2270
2275
2280
2285
2290
2295
2300
2305
2310
2315
2320
2325
2330
2335
2340
2345
2350
2355
2360
2365
2370
2375
2380
2385
2390
2395
2400
2405
2410
2415
2420
2425
2430
2435
2440
2445
2450
2455
2460
2465
2470
2475
2480
2485
2490
2495
2500
2505
2510
2515
2520
2525
2530
2535
2540
2545
2550
2555
2560
2565
2570
2575
2580
2585
2590
2595
2600
2605
2610
2615
2620
2625
2630
2635
2640
2645
2650
2655
2660
2665
2670
2675
2680
2685
2690
2695
2700
2705
2710
2715
2720
2725
2730
2735
2740
2745
2750
2755
2760
2765
2770
2775
2780
2785
2790
2795
2800
2805
2810
2815
2820
2825
2830
2835
2840
2845
2850
2855
2860
2865
2870
2875
2880
2885
2890
2895
2900
2905
2910
2915
2920
2925
2930
2935
2940
2945
2950
2955
2960
2965
2970
2975
2980
2985
2990
2995
3000
3005
3010
3015
3020
3025
3030
3035
3040
3045
3050
3055
3060
3065
3070
3075
3080
3085
3090
3095
3100
3105
3110
3115
3120
3125
3130
3135
3140
3145
3150
3155
3160
3165
3170
3175
3180
3185
3190
3195
3200
3205
3210
3215
3220
3225
3230
3235
3240
3245
3250
3255
3260
3265
3270
3275
3280
3285
3290
3295
3300
3305
3310
3315
3320
3325
3330
3335
3340
3345
3350
3355
3360
3365
3370
3375
3380
3385
3390
3395
3400
3405
3410
3415
3420
3425
3430
3435
3440
3445
3450
3455
3460
3465
3470
3475
3480
3485
3490
3495
3500
3505
3510
3515
3520
3525
3530
3535
3540
3545
3550
3555
3560
3565
3570
3575
3580
3585
3590
3595
3600
3605
3610
3615
3620
3625
3630
3635
3640
3645
3650
3655
3660
3665
3670
3675
3680
3685
3690
3695
3700
3705
3710
3715
3720
3725
3730
3735
3740
3745
3750
3755
3760
3765
3770
3775
3780
3785
3790
3795
3800
3805
3810
3815
3820
3825
3830
3835
3840
3845
3850
3855
3860
3865
3870
3875
3880
3885
3890
3895
3900
3905
3910
3915
3920
3925
3930
3935
3940
3945
3950
3955
3960
3965
3970
3975
3980
3985
3990
3995
4000
4005
4010
4015
4020
4025
4030
4035
4040
4045
4050
4055
4060
4065
4070
4075
4080
4085
4090
4095
4100
4105
4110
4115
4120
4125
4130
4135
4140
4145
4150
4155
4160
4165
4170
4175
4180
4185
4190
4195
4200
4205
4210
4215
4220
4225
4230
4235
4240
4245
4250
4255
4260
4265
4270
4275
4280
4285
4290
4295
4300
4305
4310
4315
4320
4325
4330
4335
4340
4345
4350
4355
4360
4365
4370
4375
4380
4385
4390
4395
4400
4405
4410
4415
4420
4425
4430
4435
4440
4445
4450
4455
4460
4465
4470
4475
4480
4485
4490
4495
4500
4505
4510
4515
4520
4525
4530
4535
4540
4545
4550
4555
4560
4565
4570
4575
4580
4585
4590
4595
4600
4605
4610
4615
4620
4625
4630
4635
4640
4645
4650
4655
4660
4665
4670
4675
4680
4685
4690
4695
4700
4705
4710
4715
4720
4725
4730
4735
4740
4745
4750
4755
4760
4765
4770
4775
4780
4785
4790
4795
4800
4805
4810
4815
4820
4825
4830
4835
4840
4845
4850
4855
4860
4865
4870
4875
4880
4885
4890
4895
4900
4905
4910
4915
4920
4925
4930
4935
4940
4945
4950
4955
4960
4965
4970
4975
4980
4985
4990
4995
5000
5005
5010
5015
5020
5025
5030
5035
5040
5045
5050
5055
5060
5065
5070
5075
5080
5085
5090
5095
5100
5105
5110
5115
5120
5125
5130
5135
5140
5145
5150
5155
5160
5165
5170
5175
5180
5185
5190
5195
5200
5205
5210
5215
5220
5225
5230
5235
5240
5245
5250
5255
5260
5265
5270
5275
5280
5285
5290
5295
5300
5305
5310
5315
5320
5325
5330
5335
5340
5345
5350
5355
5360
5365
5370
5375
5380
5385
5390
5395
5400
5405
5410
5415
5420
5425
5430
5435
5440
5445
5450
5455
5460
5465
5470
5475
5480
5485
5490
5495
5500
5505
5510
5515
5520
5525
5530
5535
5540
5545
5550
5555
5560
5565
5570
5575
5580
5585
5590
5595
5600
5605
5610
5615
5620
5625
5630
5635
5640
5645
5650
5655
5660
5665
5670
5675
5680
5685
5690
5695
5700
5705
5710
5715
5720
5725
5730
5735
5740
5745
5750
5755
5760
5765
5770
5775
5780
5785
5790
5795
5800
5805
5810
5815
5820
5825
5830
5835
5840
5845
5850
5855
5860
5865
5870
5875
5880
5885
5890
5895
5900
5905
5910
5915
5920
5925
5930
5935
5940
5945
5950
5955
5960
5965
5970
5975
5980
5985
5990
5995
6000
6005
6010
6015
6020
6025
6030
6035
6040
6045
6050
6055
6060
6065
6070
6075
6080
6085
6090
6095
6100
6105
6110
6115
6120
6125
6130
6135
6140
6145
6150
6155
6160
6165
6170
6175
6180
6185
6190
6195
6200
6205
6210
6215
6220
6225
6230
6235
6240
6245
6250
6255
6260
6265
6270
6275
6280
6285
6290
6295
6300
6305
6310
6315
6320
6325
6330
6335
6340
6345
6350
6355
6360
6365
6370
6375
6380
6385
6390
6395
6400
6405
6410
6415
6420
6425
6430
6435
6440
6445
6450
6455
6460
6465
6470
6475
6480
6485
6490
6495
6500
6505
6510
6515
6520
6525
6530
6535
6540
6545
6550
6555
6560
6565
6570
6575
6580
6585
6590
6595
6600
6605
6610
6615
6620
6625
6630
6635
6640
6645
6650
6655
6660
6665
6670
6675
6680
6685
6690
6695
6700
6705
6710
6715
6720
6725
6730
6735
6740
6745
6750
6755
6760
6765
6770
6775
6780
6785
6790
6795
6800
6805
6810
6815
6820
6825
6830
6835
6840
6845
6850
6855
6860
6865
6870
6875
6880
6885
6890
6895
6900
6905
6910
6915
6920
6925
6930
6935
6940
6945
6950
6955
6960
6965
6970
6975
6980
6985
6990
6995
7000
7005
7010
7015
7020
7025
7030
7035
7040
7045
7050
7055
7060
7065
7070
7075
7080
7085
7090
7095
7100
7105
7110
7115
7120
7125
7130
7135
7140
7145
7150
7155
7160
7165
7170
7175
7180
7185
7190
7195
7200
7205
7210
7215
7220
7225
7230
7235
7240
7245
7250
7255
7260
7265
7270
7275
7280
7285
7290
7295
7300
7305
7310
7315
7320
7325
7330
7335
7340
7345
7350
7355
7360
7365
7370
7375
7380
7385
7390
7395
7400
7405
7410
7415
7420
7425
7430
7435
7440
7445
7450
7455
7460
7465
7470
7475
7480
7485
7490
7495
7500
7505
7510
7515
7520
7525
7530
7535
7540
7545
7550
7555
7560
7565
7570
7575
7580
7585
7590
7595
7600
7605
7610
7615
7620
7625
7630
7635
7640
7645
7650
7655
7660
7665
7670
7675
7680
7685
7690
7695
7700
7705
7710
7715
7720
7725
7730
7735
7740
7745
7750
7755
7760
7765
7770
7775
7780
7785
7790
7795
7800
7805
7810
7815
7820
7825
7830
7835
7840
7845
7850
7855
7860
7865
7870
7875
7880
7885
7890
7895
7900
7905
7910
7915
7920
7925
7930
7935
7940
7945
7950
7955
7960
7965
7970
7975
7980
7985
7990
7995
8000
8005
8010
8015
8020
8025
8030
8035
8040
8045
8050
8055
8060
8065
8070
8075
8080
8085
8090
8095
8100
8105
8110
8115
8120
8125
8130
8135
8140
8145
8150
8155
8160
8165
8170
8175
8180
8185
8190
8195
8200
8205
8210
8215
8220
8225
8230
8235
8240
8245
8250
8255
8260
8265
8270
8275
8280
8285
8290
8295
8300
8305
8310
8315
8320
8325
8330
8335
8340
8345
8350
8355
8360
8365
8370
8375
8380
8385
8390
8395
8400
8405
8410
8415
8420
8425
8430
8435
8440
8445
8450
8455
8460
8465
8470
8475
8480
8485
8490
8495
8500
8505
8510
8515
8520
8525
8530
8535
8540
8545
8550
8555
8560
8565
8570
8575
8580
8585
8590
8595
8600
8605
8610
8615
8620
8625
8630
8635
8640
8645
8650
8655
8660
8665
8670
8675
8680
8685
8690
8695
8700
8705
8710
8715
8720
8725
8730
8735
8740
8745
8750
8755
8760
8765
8770
8775
8780
8785
8790
8795
8800
8805
8810
8815
8820
8825
8830
8835
8840
8845
8850
8855
8860
8865
8870
8875
8880
8885
8890
8895
8900
8905
8910
8915
8920
8925
8930
8935
8940
8945
8950
8955
8960
8965
8970
8975
8980
8985
8990
8995
9000
9005
9010
9015
9020
9025
9030
9035
9040
9045
9050
9055
9060
9065
9070
9075
9080
9085
9090
9095
9100
9105
91

Coding Procedure

The first step in the training of an ANN in accordance with the invention is the translation of peptide sequences into an appropriate representation. The most straightforward approach is to represent every residue by its name. However, this approach has many 5 disadvantages. First, this would result in a large input layer 2, increasing the probability of overfitting with loss of predictive ability by the ANN 1. T. Masters "Practical Neural Network Recipes in C++", Acad. Press Inc. Boston (1993). Second, the similarities of certain amino acids would be lost. For example, the relationship between leucine and either isoleucine or lysine would be treated the same. Encoding such interrelationships 10 (K. Tomii and M. Kanehisa, *Protein Eng.* 9, 27 (1996)) should increase the level of ANN generalization. Thus, a representation was chosen based upon the amino acid features presented in Tables 1A. W.R. Taylor, *J. Theor. Biol.* 119, 205 (1986). Table 1A defines 10 features associated with various amino acids (represented by standard one letter codes). Table 1B then maps each of the 20 natural amino acids as a vector of 10 binary 15 numbers, each numeric position corresponding to the feature mapping in Table 1A. A "1" indicates that the corresponding property is present. A "0" indicates that the corresponding property is absent.

8
0000214500-0000000000
0000214500-0000000000

TABLE 1A
 Clustering of amino acids according to their physico chemical features

No.	Feature	amino acid one-letter codes
0	hydrophobic	HWYFMLIVCAGTK
1	aliphatic	LIV
2	aromatic	FYWH
3	polar	TSNDEQURKHWY
4	charged	DERKH
5	positive	RKH
6	small	PVCAGTSND
7	tiny	AGS
8	glycine	G
9	proline	P

TABLE 1B
 Feature based binary coding of amino acids

amino acid	feature based code
	0123456789
G	1000001110
A	1000001100
V	1100001000
L	1100000000
I	1100000000
S	0001001100
T	1001001000
D	0001101000
N	0001001000
K	1001110000
E	0001100000
Q	0001000000
R	0001110000
H	1011110000
F	1010000000
C	1000001000
W	1011000000
Y	1011000000
M	1000000000
P	0000001001

For example, in FIGURE 1, a peptide having the amino acid sequence of "SNPSFRPFA" is coded as a binary pattern beginning with the binary pattern for "S", and continuing

with the binary pattern for “N”, *etc.* Of course, other mappings are possible, as well as other, fewer, and/or additional features.

ANN Training

As indicated in FIGURE 1, the ANN 1 has two output nodes 7a, 7b. The output signal of the ANN 1 was defined as follows:

- “00” (both nodes 7a, 7b off) denotes a non-binding sequence
- “10” (first node 7a off, second node 7b on) denotes a weakly binding sequence
- “11” (both nodes 7a, 7b on) denotes a strongly binding sequence.

The 181 K^b binding peptides were divided into strong and weak binding classes, according to their respective experimentally measured binding constants. Additionally, the 129 peptides having no detectable affinity for K^b were used as negative examples. The entire 310 peptide data base was divided into training and testing sets. In this example, the testing set contained about 1/3 of the total number of peptides. A conjugate gradient procedure (T. Masters, *“Practical Neural Network Recipes in C++”*, Acad. Press Inc. Boston (1993)) was used to determine the ANN weights, whose initial values were uniform pseudo-random numbers with a range of [-0.7, 0.7]. The network performance, defined as the mean square distance between the network output (*i.e.*, predicted binding strength) and experimentally observed value (*i.e.*, the known value of the binding strength), was measured as a function of the number of learning cycles or “epochs”. One epoch occurs when the full set of training patterns is presented to the network.

FIGURE 2 is a graph showing performance of the experimental ANN 1 on the training and testing sets as a function of training time, measured by the number of epochs. As shown in FIGURE 2, while the error in the training set decreases monotonically with an increasing number of epochs, the testing set error reaches a minimum and then slowly grows as the ANN memorizes the training set, *i.e.*, as “over fitting” occurs. T. Masters,
25 “*Practical Neural Network Recipes in C++*”, Acad. Press Inc. Boston (1993). Thus, the ANN 1 weights where chosen where the error for the test set was approximately at a

minimum. It was empirically determined that 10 hidden units were an optimal number by maximizing the performance on the testing set. Inclusion of an additional hidden layer did not change the performance in this instance.

It is expected that the relationship of the output of the ANN 1 to the experimentally determined binding constant is nonlinear. Experience is required to establish the threshold below which binding would not occur. In the preferred embodiment, the output of the ANN 1 is mapped to such empirical data as three relative classes: strongly binding, weakly binding, and nil binding.

Blind Test of the ANN

The trained ANN 1 was used to predict the binding peptides from the sequence of chicken ovalbumin, a protein containing well characterized K^b epitopes. The 11 strongest, predicted binding peptides are shown in Table 2.

TABLE 2
Comparison of Predicted Binding Peptides with Experiment Results

	Peptide	Amino Acids	ANN	K _D (moles/liter)	FACS Analysis % SIINFEKL
100X	1	SIINFEKL	0.46	3.0E-9	100
	2	SALAMVYL	0.44	7.1E-9	100
	3	AEERYPIL	0.36	6.7E-5	42
	4	NAIVFKGL	0.32	1.3E-8	76
	5	KVVRFDKL	0.27	2.6E-8	94
	6	RGDKLPGFG	0.26	5.5E-4	30
	7	DVYSFSLA	0.24	7.0E-8	65
	8	GTMSMLVL	0.23	1.2E-6	0
	9	ASEKMKIL	0.22	5.5E-4	4
	10	DHPFLFCI	0.20	4.7E-5	38
	11	ENIFYCPI	0.19	9.4E-8	77
	(VSV8)	RGYVYQGL	<i>no data</i>	4.1E-9	<i>not applicable</i>

Following are explanations of each column:

Peptide. Peptides 1-11 are from the ovalbumin sequence listed in order predicted by the ANN 1 to bind K^b. VSV8 is the peptide epitope from vesicular stomatitis virus nucleoprotein used as the reporter peptide in competition binding assays (see discussion of FIGURE 3 below).

5 **ANN.** Relative binding strengths predicted by the ANN 1 defined as the value of the output signal on the second node 7b of the output layer 6. For all sequences presented here, the output value of the first node 7a is 0.7 (the threshold value).

10 **K_D.** Dissociation constants of the predicted peptides, in moles/liter. Dissociation curves used to predict the K_D values for peptides 2-11 are shown in FIGURE 3 . Peptide 1 is the known immunodominant epitope for ovalbumin and has been characterized previously.

15 **FACS Analysis.** Values from fluorescence activated cell sorter (FACS) analysis showing the relative amounts of K^b on the surface of K^b transfected drosophila cells following an 18-hour incubation with the indicated peptides. Cells were strained with the anti mouse MHC class 1 antibody Y3 followed by a fluoresceine conjugated second antibody. Median fluorescence values from separate experiments were normalized by subtracting the median fluorescence obtained in the absence of added peptides from each peptide sample and then expressing those values as the percent of the fluorescence obtained with SIINFEKL (which was examined in all experiments).

Validation of ANN Predictions

20 To experimentally test the predictions, these 11 peptides were synthesized. Experimental binding affinities for K^b were determined by a competition assay previously used to determine the dissociation constants of peptides for mouse class I molecules. M. Matsumura, Y. Saito, M.R. Jackson, E.S. Song and P.A. Peterson, *J. Biol. Chem.* 267(33), 23589 (1992); Y. Saito, P.A. Peterson and M. Matsumura, *J. Biol. Chem.* 268(28), 21309 (1993); R. Miller, *Methods Enzymology*, 92, 589 (1983).

FIGURE 2 is a graph showing the competition binding assay for 11 peptides under test. VSV8 (see Table 2) was radio-iodinated (chloramine-T) for use as a tracer peptide. Competitor peptides 2-11 are ANN predicted K^b binding peptides added to 100,000 cpm of the tracer peptide ($2.1 \times 10^4 \mu\text{M}$) with concentrations of K^b that, in the absence of 5 competitors, bound about half of the added tracer. The graph shows the concentration dependent inhibition of the tracer peptide binding by the added competitor peptides. The curve labeled VSV8 are the results of a control experiment where the competitor peptide was the same as the tracer. Peptide concentrations are in moles/liter.

Referring again to Table 2, specific peptide epitopes bind to K^b having K_D values below 10 10^{-7} M . Of the first seven peptides predicted to bind the strongest, five bound at biologically significant levels. This translates into a hit rate of slightly better than 70%. For those peptides that bound strongly, their affinities were predicted in the same order as determined experimentally. The other two peptides in this group bound at levels with lesser or equal affinities to the average K_D ($40 \mu\text{M}$).

15 In agreement with the experimental analysis, the top two predicted peptides were in fact the strongest binders and included the immunodominant epitope, OVA-8, for K^b . This result is significant as there are 20 peptides in the ovalbumin sequence which contain internal anchor residues, and the ANN analysis narrowed this field to one, OVA-8. Moreover, the second best binding peptide contains no anchor amino acids in positions 20 three or five, and thus would not have been predicted using a simple statistical analysis.

Peptide binding was also analyzed by the ability to stabilize cell surface K^b molecules. Empty class I molecules are thermolabile, but they can be stabilized by binding appropriate peptides. Peptides were bound to K^b molecules expressed on the surfaces of 25 K^b transfected Drosophila cells. Their relative binding strengths are indicated by their median fluorescence. As shown in Table 2, at 23°C , the ability of the peptides to stabilize K^b closely mirrored their binding affinities determined by the competition assay.

RECEIVED
JULY 22 1990
U.S. GOVERNMENT
PRINTING OFFICE
1990 500-100-00000

Summary

A list of 30 binding peptides were predicted along with scores for the predicted relative binding affinities. To evaluate these predictions, the 11 peptides at the top of the list were synthesized and their binding affinities determined experimentally. Our results
5 demonstrate that the ANN 1 can make highly accurate predictions, some of which could not have been predicted manually using extant anchor position based binding rules. Five of the predicted seven best binders bound with good affinity ($K_D < 10^{-7}$ nM). Most significantly, the top predicted peptide bound the strongest and is the known immuno-dominant epitope. Furthermore, despite the fact that the second best predicted peptide
10 lacked internal anchor residues and thus would not have been included in the set of 20 manually predicted sequences, it was shown experimentally to bind with the second strongest affinity. This affinity is greater than four other predicted binding peptides in the top eleven scores, which do contain internal anchor residues.

Two peptides in the top 7 did not bind K^b with significant affinity; the question is why.
15 One possibility is that binding to phage somehow does not accurately simulate peptide binding in all cases. Other possible reasons for these nonbinding sequences are that an insufficiently diverse combination of amino acids was present in the positive and negatively selected phage sequences or that the system of encoding amino acids for the ANN did not adequately distinguish the chemical and physical properties of all of the
20 amino acids. These alternatives are presently being analyzed to improve accuracy of the invention. However, the success rate in the top seven predictions shows that the ANN approach works well.

In its present application, the ANN analysis should be able to predict class I binding peptides for an unlimited number of protein antigens. This may further the understanding
25 of the class I molecular structure as it pertains to peptide binding and perhaps further elucidate how these binding interactions pertain to function. More generally, the inventive approach represents but a first application for identifying binding motifs from either peptide or even small molecule (e.g., peptide mimetics) combinatorial libraries. One

14

strength of the invention is that it allows one to generalize and extract the latent information encoded in a random peptide library that has been screened for a particular property or functionality. The results of applying the ANN 1 of the invention may be used to design stronger binding sequences.

5 *Implementation*

The ANN 1 of the invention may be implemented in hardware or software, or a combination of both. However, preferably, the invention is implemented in computer programs executing on programmable computers each comprising at least one processor, at least one data storage system (including volatile and non-volatile memory and/or storage elements), at least one input device, and at least one output device. Program code is applied to input data to perform the functions described herein and generate output information. The output information is applied to one or more output devices, in known fashion.

10 Each program is preferably implemented in a high level procedural or object oriented programming language to communicate with a computer system. However, the programs can be implemented in assembly or machine language, if desired. In any case, the language may be a compiled or interpreted language.

15 Each such computer program is preferably stored on a storage media or device (e.g., ROM or magnetic diskette) readable by a general or special purpose programmable computer, for configuring and operating the computer when the storage media or device is read by the computer to perform the procedures described herein. The inventive system may also be considered to be implemented as a computer-readable storage medium, configured with a computer program, where the storage medium so configured causes a computer to operate in a specific and predefined manner to perform the functions 20 described herein.

25

DECEMBER 2000
2000-09-26 09:00:00

/S/

A number of embodiments of the present invention have been described. Nevertheless, it will be understood that various modifications may be made without departing from the spirit and scope of the invention. Accordingly, it is to be understood that the invention is not to be limited by the specific illustrated embodiment, but only by the scope of the appended claims.

0 8 8 6 2 1 9 2 " 0 6 2 3 3 7